RNA-Seq QC Workflow

User Documentation and Tutorial

CCRIFX Bioinformatics Core
Date: Oct 31, 2013

Purpose

This workflow is used to automate the generation of a series of quality control metrics for RNA-Seq data using BAM files as input. Various Picard (v1.6.1) and RNA-SeQC (v1.1.7) alignment metrics are computed from BAM files and user-friendly aggregated reports are compiled from the generated metrics. It supports integrated and automated pre-processing of the BAM files to make them suitable for the QC modules to work correctly.

Introduction

For RNA-Seq data analysis, it is valuable to compare sequencing quality across different samples or experiments and evaluate different experimental parameters, before continuing with downstream analysis. Both the Picard and RNA-SeQC tools from Broad Institute have various scripts to assess different metrics from input alignment files.

The metrics that are generated by this workflow are:

- Picard CollectRnaSeqMetrics : Program to collect metrics about the alignment of RNA to various functional classes of loci in the genome: coding, intronic, UTR, intergenic, ribosomal.
- Picard CollectMultipleMetrics : This command can be run with one or more of the following options: CollectAlignmentSummaryMetrics, CollectInsertSizeMetrics, QualityScoreDistribution, MeanQualityByCycle
- Picard MarkDuplicates: Examines aligned records in the supplied SAM or BAM file to locate duplicate molecules. All records are then written to the output file with the duplicate records flagged.
- RNA-SeQC metrics: Generates HTML reports and tab delimited files of metrics data

This workflow comprises of two scripts which together act as a wrapper to automate the parallel execution of routines for generation of various metrics for multiple samples. The two scripts are: a) RNASeqQC.09.14.2012.pl and b) RNASeqQCReport.09.06.2012.pl that have to be run in succession.

Briefly, an input BAM file is passed to the script RNASeqQC.09.14.2012.pl. This script first pre-processees to make it suitable for running the scripts. These pre-processing operations are:
1) Clean sam (samtools 1.1.18 and picard 1.61)
2) AddReplaceReadGroups (picard 1.61)
3) Make a sequence dictionary (picard 1.61)
4) ReOrder Bam (picard 1.61)
5) Optionally use the filter workflow EmptyReadBamFilter.09.19.2012 to remove reads marked with a "*" in column 11 of BAM file
6) ReOrder (a second time for CollectRNASeqMetrics) (picard 1.61)
7) Mergesort (samtools 1.1.18)
8) Index (samtools 1.1.18)

The final file appears in a filed called "prepped.bam" along with its index "prepped.bam.bai' in its tmpdir directory.

The final pre-processed BAM file is then used to generate various metrics based on a) Picard (CollectRNASeqMetrics.jar, CollectMultipleMetrics.jar, MarkDuplicates.jar) and b) RNA-SeQC metrics. Internally the CollectMultipleMetrics is passed the following metrics programs to generate various metrics (CollectAlignmentSummaryMetrics, CollectInsertSizeMetrics, QualityScoreDistribution, MeanQualityByCycle).

The output from RNASeqQC.09.14.2012.pl (from individual sample directories) is then used by RNASeqQCReport.09.06.2012.pl to aggregate the

metrics across samples and generate formatted text files that can be readily imported into excel for review/interpretation.

Prerequisites

This documentation assumes that the user has either a Biowulf or Moab account. The tutorial provided below assumes that execution is carried out on Biowulf
Picard tools and RNA-SeQC are required to be installed on the system
The reference FASTA file (and the index) used to generate the BAM file has to exist on the system. A GTF file (whose contents are taken as reference genomic data) needs to be passed as a full-path to the program.
For generating rRNA metrics, a full path to an rRNA intervals file is required.
During the execution of multimetrics.pl, R (version 2.15) is used for PDF generation.

Input

BAM file directory or file (--basebam) -the path in whose contents are searched recursively for BAM files and whose contents will include the script output after successful execution. REQUIRED. Should be a full path.

REFFLAT file (--refflat) - a full path to a refflat file to be used with the QC programs. A gzip-compressed reference file is usable. An uncompressed file is also okay.

Reference FASTA (--fasta) - The fasta file (whose contents are taken as reference genomic data) passed to the analysis programs. REQUIRED. Should be a full path

GTF file (--gtf) - The GTF file (whose contents are taken as reference annotation data) passed to the analysis programs (RNASEQQC, etc) REQUIRED. Should be a full path

Option to merge BAM files or process individually (--procbamind) - process bam files individually, instead of merging them. In this case, the output directory has additional levels of structure where directories with serial numbers correspond to fasta files found for processing. In this case, the program calls itself recursively with queuemode set to 'y'. REQUIRED. Should be either 'y' or 'n' (case insensitive)

Queue-mode (--queuemode) - cause the script to generate a job file of itself and to submit itself. REQUIRED Should be either 'y' or 'n' (case insensitive).

Output directory (--outdir) - An output directory. It must NOT exist before the script is run. The script creates it. This is to avoid file namespace collisions. If the script is unable to create the directory, then the script will fail. REQUIRED

rRNA intervals file (--rrnaintervals) -A full path to an rRNA intervals file used by collectrnaseqmetrics

Whether to search only for "accepted_hits.bam" (--acchit)- Search ONLY for BAM files named "accepted_hits.bam" REQUIRED. Should be either 'y' or 'n' (case-insensitive) . If 'n' then all .bam files will be used in QC.

Output

Metrics generated by RNASeqQC.09.14.2012.pl and aggregated by RNASeqQCReport.09.06.2012.pl

The search directory is scanned for metrics.tsv files (from RNA-SeQC), for alignment_metrics_summary files and insert_metrics_summary files (from Picard). The output directory is created. A new metrics.tsv file is created with aggregated data from the individual metrics.tsv files found. Similarly aggregated files for the alignment and insert metrics files are also created. They are named alignment_metrics_summary and insert_metrics_summary respectively. "Formatted" files are created of the insert_metrics output and of the metrics.tsv output. The have the same information but with columns reordered and/or sub-set for easier reading/interpretation.

Metrics generated by RNASeqQC.09.14.2012.pl but NOT aggregated by RNASeqQCReport.09.06.2012.pl

There are various other metrics, reports and graphs that are generated by RNASeqQC, but are not being parsed by RNASeqQCReport for reporting purposes. As of now, the user is encouraged to manually review and compile these additional metrics if interested.

Tutorial

In the tutorial section, the unix commands are highlighted in **bold font**. The standard output, resulting from issuing unix commands, is highlighted in red font. The dataset used in the tutorial contains eight BAM files generated by the rnapipe workflow. The BAM files are located in the following directory on Biowulf:
data/CCRIFX/TUTORIAL/QC_Jul_30_2013/rnapipe/.

1. Setting up environments

a. Use scripts from the latest release of the workflow scripts (release folders have the yyyy-mm-dd naming convention, pick the one with the most recent date). In this case, we have the latest release dated 2013-08-29

(/data/CCRIFX/RELEASE_08.29.2013/ccrifx_root_committed/bin/)

b. For general information on setting up environments, type more /data/CCRIFX/RELEASES/README.txt. Then browse the following directory /data/CCRIFX/RELEASE_08.29.2013/ccrifx_root_committed/

Read through the following documentation to familiarize yourself with content and what to expect: release_notes.txt, QUICK_START.txt and WF_EXAMPLE.txt.

c. Add the environment to your path using the following command (again, replace the release folder with the most recent release at the time of execution)

> **source /data/CCRIFX/RELEASE_08.29.2013/ccrifx_root_committed/ccrifx_sys_setup.sh**

d. After typing the above command, the $PATH environment variable is updated. You can check for this using the following unix commands (**echo $PATH**). The environment variable $CCRIFX_SYS_ROOT should also be defined. Check the environment variable $CCRIFX_SYS_ROOT is set by using the following unix command (**ls $CCRIFX_SYS_ROOT**). One of the benefits of setting these environments is so that you have access to the scripts plus dependencies without having to refer to them via full path (see step 2). If any error message is returned whilst setting environments – please report the issue using the issue-reporting template.

Review the standard output of the source command. The standard out will provide a list of third party software (ie dependencies of the scripts and includes information regarding the versions of bowtie, cufflinks, cuffdiff and cuffmerge installed on biowulf). To find out what versions of software are available to you on biowulf refer to reference [2]. You will need to know what version of the software from the RNASeqQC pipeline that you wish the run.

2. Run RNASeqQC pipeline using the script RNASeqQC.09.14.2012.pl

Here are some of the decisions to be made in order to set the values for the parameter options for the main RNASeqQC.09.14.2012.pl script.

a) –filter: Apply the "emptyread" filter workflow as the final step of data preparation before RNASEQQC is called (default y) REQUIRED. Should be 'y' or 'n' (case-insensitive). Setting this value to 'y' invokes the scrip EmptyReadBamFilter.09.19.2012.pl internally, in order to filter our mappings with "*" for quality in Column 11 of the BAM files. The essence of this script is to filter out "bad" reads from a BAM file and write them to a SAM file. A read is bad if columns 11 is a "*". The stats txt file informs the user of the number of reads read from the bam file and the number of reads written to the sam file. The numbers can be used to give an indication of how many reads were filtered out.

b) –basebam: The full path to the directory containing the BAM files. All your BAM files need not be in a single directory, because internally the script 'finds' all .BAM files recursively under the base BAM directory. –refflat

c) –procbamind: If set to 'y', the script will process BAM files individually. If set to 'n' then the BAM files will be merged. In this case, the output directory has additional levels of structure where directories with serial numbers correspond to fasta files found for processing. In this case, the program calls itself recursively with queuemode set to 'y'. REQUIRED Should be either 'y' or

'n' (case insensitive). In this case, each BAM file belongs to one sample and hence we set this flag to ‘y’.

d) –refflat: This is the refFlat.txt.gz file from the reference genome. In this case, we will use the Ensembl: Homo_sapiens refFlat file that is located at: /fdb/igenomes/Homo_sapiens/UCSC/hg19/Annotation/Genes/refFlat.txt.gz

e) –gtf: The annotation GTF file. In this case we will use /fdb/igenomes/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf

f) –fasta: This is the reference FASTA. Here will will use: /fdb/igenomes/Homo_sapiens/ UCSC/hg19/Sequence/WholeGenomeFasta/genome.fa

g) –queuemode: cases the script to generate a job file of itself and to submit itself. Here we set it to ‘y’ as we want automated job submission

h) –outdir: An output directory. It must NOT exist before the script is run. The script creates it. This is to avoid file namespace collisions. If the script is unable to create the directory, then the script will FAIL. Here we use this path: /data/CCRIFX/TUTORIAL/QC_Jul_30_2013/rnapipeOUT7/RNASeqQC_output

–acchit: Search only for BAM files named “accepted_hits.bam”. This is very important. Usually we set this to ‘y’ as we do not want the script to generate QC metrics on intermittent BAM files generated by Tophat. Here, we set this to ‘y’.

j) –rrnaintervals: A file containing the genomic coordinates of rRNA. This file is in GATK format and uses the .list extension. The file contains one genomic coordinate per line in the following format:
chr:start-stop

If this file is not provided, the information is drawn from the annotation GTF file. A full path to the rRNA intervals files used by CollectRNASeqMetrics. This file is not publicly distributed and hence has to be created using the Picard IntervalListTools (http://picard.sourceforge.net/command-line-overview.shtml#IntervalListTools).
Here we use the path:
/data/CCRIFX/rRNAIntervals/fdb/igenomes/Homo_sapiens/UCSC/hg19/Annotation/Genes/rRNA.intervals.formatted.txt

The command used to run the script with all the options above is:

> **RNASeqQC.09.14.2012.pl --filter y --basebam /data/CCRIFX/TUTORIAL/QC_Jul_30_2013/rnapipeOUT7 --refflat /fdb/igenomes/Homo_sapiens/UCSC/hg19/Annotation/Genes/refFlat.txt.gz --gtf /fdb/igenomes/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf --fasta /fdb/igenomes/Homo_sapiens/UCSC/hg19/Sequence/WholeGenomeFasta/genome.fa --queuemode y --procbamind y --outdir /data/CCRIFX/TUTORIAL_rnaseqqc_out/ --acchit y --rrnaintervals /data/CCRIFX/rRNAIntervals/fdb/igenomes/Homo_sapiens/Ensembl/GRCh37/Annotation/Genes/rRNA.intervals.formatted.txt**

On submitting this command, you will see the following output on the terminal:

mkdir: created directory `/data/CCRIFX/TUTORIAL_rnaseqqc_out'
4586293

Now list the contents of the output directory: /data/CCRIFX/TUTORIAL_rnaseqqc_out. You should see the following files created by the script:

jailwalapa@biowulf TUTORIAL_rnaseqqc_out $ ls -al

```
total 56
drwxr-s---      3       jailwalapa      CCRIFX          4096 Oct 29 10:10 .
drwxrws---     68       stahlbergea     CCRIFX          8192 Oct 29 10:10 ..
-rw-------      1       jailwalapa      jailwalapa     18569 Oct 29 10:10 4586293.biobos.OU
-rw-r-----      1       jailwalapa      CCRIFX          1614 Oct 29 10:10 RNASEQQC.job
drwxr-s---     10       jailwalapa      CCRIFX          4096 Oct 29 10:10
RNASeqQC.09.14.2012_out
-rw-r-----      1       jailwalapa      CCRIFX         15006 Oct 29 10:10 rnaseqqclog.txt
```

Now check the queue for the automated job submissions. You should see 8 jobs either queued or running on 70gb nodes:

jailwalapa@biowulf TUTORIAL_rnaseqqc_out $ qstat -u jailwalapa

```
biobos:
                                                  Req'd  Req'd   Elap
Job ID          Username Queue    Jobname    SessID NDS TSK Memory Time  S Time
--------------- -------- -------- ---------- ------ --- --- ------ ----- - -----
4586294.biobos  jailwala g72      BIOWULF     6234   1   1   70gb   --  R 00:01
4586295.biobos  jailwala g72      BIOWULF    28585   1   1   70gb   --  R 00:01
4586296.biobos  jailwala g72      BIOWULF    16068   1   1   70gb   --  R 00:01
4586297.biobos  jailwala g72      BIOWULF    23942   1   1   70gb   --  R 00:01
4586298.biobos  jailwala g72      BIOWULF    17568   1   1   70gb   --  R 00:01
4586299.biobos  jailwala g72      BIOWULF     5831   1   1   70gb   --  R 00:01
4586300.biobos  jailwala g72      BIOWULF    12304   1   1   70gb   --  R 00:01
4586301.biobos  jailwala g72      BIOWULF     4475   1   1   70gb   --  R 00:01
```

So, while you are waiting for these jobs to complete, here is a description of what the script is doing internally. First, the input directory (BAM_DIR) is

scanned with the find command. Using samtools (merge v0.1.18), any encountered bam files are merged (into an intermediate bam file). Second, once merged (samtools 0.1.18), the bam is passed through RNASEQQC.Prep (9.11.2012). The prepped bam and its index are then retrieved from the PREP area and submited to RNASEQQC (v1.1.7), Picard CollectRNASeq Metrics, and Collect MultipleMetrics (insert and alignment) are run next.. All of these picard tools are version 1.61. Finally, picard MarkDuplicates v1.61 is run.
Files generated by multimetrics: MultiMet.insert_size_histogram.pdf, MultiMet.insert_size_metrics, MultiMet.quality_by_cycle.pdf, MultiMet.quality_by_cycle_metrics, MultiMet.quality_distribution.pdf, MultiMet.quality_distribution_metrics, and MultiMet.alignment_summary_metrics.

A script written and used to call the picard programs (collect multiple metrics, collect rna seq metrics, mark duplicates): multimet.sh. Note that during the execution of multimetrics R (version 2.15) is used for PDF generation.

The sample names are picked up from the fastq file names when the bam files are scanned. The RNASEQQC script looks inside each BAM file. It looks for the @PG line that has the mapping command and fastq input files. The fastq file names are found there and that is where the sample names are extracted from by the workflow. If no fastq file name is found there, then a "default" sample name is used. If you use the QC workflow for catting, then the sample names *should* be in the regular expression it uses. The RNASEQQC uses the same regular expression/format to extract the sample names.

3. Check the output folder for the results from RNASeqQC.09.14.2012.pl

Once you get email notifications about jobs completions, issue the qstat command once again to ensure that jobs for all the samples are complete.

Here is what directory structure and files are generated. For example, lets take a look at the directory structure for one of the samples (4MRT):

jailwalapa@biowulf TUTORIAL_rnaseqqc_out $ tree

```
|-- 4586293.biobos.OU
|-- RNASEQQC.job
|-- RNASeqQC.09.14.2012_out
|  |-- RNASEQ_0_4MRT
|  |  |-- RNASeq
|  |  |  |-- 4586294.biobos.OU
|  |  |  |-- RNASEQQC.job
|  |  |  |-- RNASeqQC.09.14.2012_out
|  |  |  |  |-- MultiMet.alignment_summary_metrics
|  |  |  |  |-- MultiMet.insert_size_histogram.pdf
|  |  |  |  |-- MultiMet.insert_size_metrics
|  |  |  |  |-- MultiMet.quality_by_cycle.pdf
|  |  |  |  |-- MultiMet.quality_by_cycle_metrics
|  |  |  |  |-- MultiMet.quality_distribution.pdf
|  |  |  |  |-- MultiMet.quality_distribution_metrics
|  |  |  |  |-- PrepTmpDir
|  |  |  |  |  |-- filter.stat
|  |  |  |  |  `-- input.bam ->
/data/CCRIFX/TUTORIAL_rnaseqqc_out//RNASeqQC.09.14.2012_out//RNASEQ_0_4MRT//RNASeq//RNASeqQC.09.14.2012_out//merged.bam
|  |  |  |  |-- RNASeq_QC
|  |  |  |  |  |-- 4MRT
|  |  |  |  |  |  |-- 4MRT.libraryComplexity.txt
|  |  |  |  |  |  |-- 4MRT.metrics.tmp.txt
|  |  |  |  |  |  |-- 4MRT.metrics.tmp.txt.chimericPairs.txt
|  |  |  |  |  |  |-- 4MRT.metrics.tmp.txt.intronReport.txt
|  |  |  |  |  |  |-- 4MRT.metrics.tmp.txt.intronReport.txt_exonOnly.txt
|  |  |  |  |  |  |-- 4MRT.metrics.tmp.txt.intronReport.txt_intronOnly.txt
|  |  |  |  |  |  |-- 4MRT.metrics.tmp.txt.introns.rpkm.gct
```

```
|  |  |  |  |  |  |-- 4MRT.metrics.tmp.txt.rpkm.gct
|  |  |  |  |  |  |-- 4MRT.metrics.txt
|  |  |  |  |  |  |-- 4MRT.rRNA_counts.txt
|  |  |  |  |  |  |-- highexpr
|  |  |  |  |  |  |  |-- 4MRT.DoCTranscripts
|  |  |  |  |  |  |  |-- 4MRT.DoCTranscriptsSummary
|  |  |  |  |  |  |  |-- 4MRT.transcripts.list
|  |  |  |  |  |  |  |-- gapLengthHistogram.txt
|  |  |  |  |  |  |  |-- index.html
|  |  |  |  |  |  |  |-- intervals.list
|  |  |  |  |  |  |  |-- meanCovByPosition.txt
|  |  |  |  |  |  |  |-- meanCovByPositionNormed.txt
|  |  |  |  |  |  |  |-- perBaseDoC.out
|  |  |  |  |  |  |  |-- perBaseDoC.out.sample_cumulative_coverage_counts
|  |  |  |  |  |  |  |-- perBaseDoC.out.sample_cumulative_coverage_proportions
|  |  |  |  |  |  |  |-- perBaseDoC.out.sample_interval_statistics
|  |  |  |  |  |  |  |-- perBaseDoC.out.sample_interval_summary
|  |  |  |  |  |  |  |-- perBaseDoC.out.sample_statistics
|  |  |  |  |  |  |  `-- perBaseDoC.out.sample_summary
|  |  |  |  |  |  |-- lowexpr
|  |  |  |  |  |  |  |-- 4MRT.DoCTranscripts
|  |  |  |  |  |  |  |-- 4MRT.DoCTranscriptsSummary
|  |  |  |  |  |  |  |-- 4MRT.transcripts.list
|  |  |  |  |  |  |  |-- gapLengthHistogram.txt
|  |  |  |  |  |  |  |-- index.html
|  |  |  |  |  |  |  |-- intervals.list
|  |  |  |  |  |  |  |-- meanCovByPosition.txt
|  |  |  |  |  |  |  |-- meanCovByPositionNormed.txt
|  |  |  |  |  |  |  |-- perBaseDoC.out
|  |  |  |  |  |  |  |-- perBaseDoC.out.sample_cumulative_coverage_counts
|  |  |  |  |  |  |  |-- perBaseDoC.out.sample_cumulative_coverage_proportions
|  |  |  |  |  |  |  |-- perBaseDoC.out.sample_interval_statistics
|  |  |  |  |  |  |  |-- perBaseDoC.out.sample_interval_summary
|  |  |  |  |  |  |  |-- perBaseDoC.out.sample_statistics
|  |  |  |  |  |  |  `-- perBaseDoC.out.sample_summary
|  |  |  |  |  |  `-- medexpr
```

```
| | | | | |     |-- 4MRT.DoCTranscripts
| | | | | |     |-- 4MRT.DoCTranscriptsSummary
| | | | | |     |-- 4MRT.transcripts.list
| | | | | |     |-- gapLengthHistogram.txt
| | | | | |     |-- index.html
| | | | | |     |-- intervals.list
| | | | | |     |-- meanCovByPosition.txt
| | | | | |     |-- meanCovByPositionNormed.txt
| | | | | |     |-- perBaseDoC.out
| | | | | |     |-- perBaseDoC.out.sample_cumulative_coverage_counts
| | | | | |     |-- perBaseDoC.out.sample_cumulative_coverage_proportions
| | | | | |     |-- perBaseDoC.out.sample_interval_statistics
| | | | | |     |-- perBaseDoC.out.sample_interval_summary
| | | | | |     |-- perBaseDoC.out.sample_statistics
| | | | | |     `-- perBaseDoC.out.sample_summary
| | | | | |-- countMetrics.html
| | | | | |-- exons.rpkm.gct
| | | | | |-- gapLengthHist_high.png
| | | | | |-- gapLengthHist_high.txt
| | | | | |-- gapLengthHist_low.png
| | | | | |-- gapLengthHist_low.txt
| | | | | |-- gapLengthHist_medium.png
| | | | | |-- gapLengthHist_medium.txt
| | | | | |-- index.html
| | | | | |-- meanCoverageNorm_high.png
| | | | | |-- meanCoverageNorm_high.txt
| | | | | |-- meanCoverageNorm_low.png
| | | | | |-- meanCoverageNorm_low.txt
| | | | | |-- meanCoverageNorm_medium.png
| | | | | |-- meanCoverageNorm_medium.txt
| | | | | |-- meanCoverage_high.png
| | | | | |-- meanCoverage_high.txt
| | | | | |-- meanCoverage_low.png
| | | | | |-- meanCoverage_low.txt
| | | | | |-- meanCoverage_medium.png
| | | | | |-- meanCoverage_medium.txt
```

```
| | | | | |-- metrics.tsv
| | | | | |-- rRNA_intervals.list
| | | | | |-- refGene.txt
| | | | | |-- refGene.txt.idx
| | | | | `-- report.html
| | | | |-- dup.metrics.txt
| | | | |-- merged.bam ->
/data/CCRIFX/TUTORIAL_rnaseqqc_out//RNASeqQC.09.14.2012_out//RNASEQ_0_4MRT/accepted_hits.bam
| | | | |-- multimet.sh
| | | | |-- prepped.bam
| | | | |-- prepped.bam.bai
| | | | `-- prepped.bam.noDups.bam
| | | `-- rnaseqqclog.txt
| | `-- accepted_hits.bam ->
/data/CCRIFX/TUTORIAL/QC_Jul_30_2013/rnapipeOUT7/4MRT_000000_tophatout/accepted_hits.bam
```

4. Run RNASeqQC report using the script RNASeqQCReport.09.06.2012.pl

Once the workflow RNASeqQC.09.14.2012.pl has executed on all samples, you can generate summary reports across all samples, with the RNASeqQCReport.09.06.2012.pl script. This is a simple script, just requiring the input directory (containing the output of RNASeqQC.09.14.2012.pl) and an output directory to hold the reports.

jailwalapa@biowulf TUTORIAL_rnaseqqc_out $ RNASeqQCReport.09.06.2012.pl --dir $PWD --out $PWD/RNASeqQCReports

```
mkdir: created directory `/data/CCRIFX/TUTORIAL_rnaseqqc_out/RNASeqQCReports'
##########################
# Thu Oct 31 09:15:42 EDT 2013
# Finding metrics files...
# find /data/CCRIFX/TUTORIAL_rnaseqqc_out -type f | grep -P 'metrics\.tsv$'|sort
```

```
##############################################
# Thu Oct 31 09:15:42 EDT 2013
# Finding CollectRNASeqSummaryMetrics files....
# find /data/CCRIFX/TUTORIAL_rnaseqqc_out | grep -P 'rnaseq_summary_metrics$'|sort

########################################
# Thu Oct 31 09:15:42 EDT 2013
# Finding alignment_summary_metrics ...
# find /data/CCRIFX/TUTORIAL_rnaseqqc_out | grep -P
'alignment_summary_metrics$'|sort

################################
# Thu Oct 31 09:15:43 EDT 2013
# Finding insert size metrics...
# find /data/CCRIFX/TUTORIAL_rnaseqqc_out | grep -P 'insert_size_metrics$'|sort
```

The input directory is scanned for metrics.tsv files (generated from RNASEQ-QC), for alignment_metrics_summary files (from picard), and for insert_metrics_summary files (from picard). The output directory is created. A new metrics.tsv file is created with aggregated data from the individual metrics.tsv files found. Similary aggregated files for the alignment and insert metrics files are also created. They are named alignment_metrics_summary and insert_metrics_summary respectively. "Formatted" files are created of the insert_metrics output and of the metrics.tsv output. The have the same information but with columns reordered and/or subsetted for easie reading/interpretation.

jailwalapa@biowulf RNASeqQCReports $ ls -al

total 36

```
drwxr-s--- 2 jailwalapa CCRIFX 4096 Oct 31 09:15 .
drwxr-s--- 4 jailwalapa CCRIFX 4096 Oct 31 09:15 ..
```

-rw-r----- 1 jailwalapa CCRIFX 4288 Oct 31 09:15 alignment_metrics_summary
-rw-r----- 1 jailwalapa CCRIFX 1195 Oct 31 09:15 insert_metrics_summary
-rw-r----- 1 jailwalapa CCRIFX 613 Oct 31 09:15 insert_metrics_summary.formatted
-rw-r----- 1 jailwalapa CCRIFX 3745 Oct 31 09:15 metrics.formatted.tsv
-rw-r----- 1 jailwalapa CCRIFX 3664 Oct 31 09:15 metrics.formatted.tsv.humanReadable.txt
-rw-r----- 1 jailwalapa CCRIFX 3544 Oct 31 09:15 metrics.tsv

Each of these files (metrics.formatted.tsv.humanReadable.txt, insert_metrics_summary.formatted and alignment_metrics_summary) can be opened in Excel and can be formatted further for creating tables that can be shared with the investigators.

References

RNA-SeQC publication

Deluca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G. (2012) RNA-SeQC: RNA-Seq metrics for quality control and process optimization. *Bioinformatics* (2012) 28 (11): 1530-1532

RNA-SeQC documentation

http://www.broadinstitute.org/cancer/cga/rna-seqc

Biowulf documentation for monitoring jobs on the queue

http://biowulf.nih.gov/user_guide.html#monitor

Unix cheat sheets

http://www.cyberciti.biz/tips/linux-unix-commands-cheat-sheets.html

Biowulf documentation on Picard

http://biowulf.nih.gov/apps/picard.html

Biowulf documentation on RNA-SeQC

http://biowulf.nih.gov/apps/rnaseqc.html

FAQ

1) How does the workflow assign sample names to each BAM file ?

The sample names are picked up from the fastq file names when the bam files are scanned. The script **RNASeqQC.09.14.2012.pl** looks inside each BAM file. It looks for the @PG line that has the mapping command and fastq input files. The fastq file names are found there and that is where the sample names are extracted from by the workflow. If no fastq file name is found there, then a "default" sample name is used. If you use the QC workflow for catting, then the sample names should be in the regular expression it uses. The **RNASeqQC.09.14.2012.pl** uses the same regular expression/format to extract the sample names.

2) Under what path on Biowulf are the reference datasets that should work with this workflow?

The RNA-SeQC program requires the reference fasta to be accompanied by an index (use samtools faidx to make one). Its execution will fail without it. The index file should be in the same directory as the fasta file, but with the name X.fai (where X is the filename of the reference fasta). Also, the reference fasta must have a dictionary (.dict) file associated with it.
The biowulf cluster filesystem has convenient datasets in the "igenomes" area (located under /fdb/igenomes)
The three files below exemplify the proper directory/filesystem arrangement of a reference fasta and accompanying .fai and .dict files:

- /fdb/igenomes/Drosophila_melanogaster/UCSC/dm3/Sequence/WholeGenomeFasta/genome.dict
- /fdb/igenomes/Drosophila_melanogaster/UCSC/dm3/Sequence/WholeGenomeFasta/genome.fa.fai
- /fdb/igenomes/Drosophila_melanogaster/UCSC/dm3/Sequence/WholeGenomeFasta/genome.fa

Contact

If you identify areas where you would like to expand the user documentation or would like to make an update to the user documentation or provide comments/feedback please contact Yvonne Edwards or Parthav Jailwala.

CCRIFX Bioinformatics Core
Advanced Biomedical Computing Center (ABCC)
Information Systems Program
Leidos Biomedical Research, Inc.
Frederick National Laboratory for Cancer Research (FNLCR)
P. O. Box B, Frederick, MD 21702
Phone: 301.594.1395
Fax: 301.480.0391
http://ccrifx.cancer.gov

Citation

If you want to cite this workflow, please use the following link:

RNA-Seq QC pipeline:
http://ccrifx.cancer.gov/apps/site/workflows_for_bioinformatics_analysis

This document was lasted updated December 13, 2013.